Early predictive model for COVID-19 mortality and ventilator support: reducing risks for patients

Anshu SS Kotia, Jaya Gupta, Vinod Singh K.

1. Introduction

Predictive models for COVID-19 mortality have been developed and kept up to date ever since the start of the epidemic. It’s essential to remember that these models are prone to change when fresh data becomes available since they were developed using the data and hypotheses that were accessible from the point in time (Alaa et al 2020). Any prediction model’s accuracy is influenced by the caliber of the data it uses and the underlying assumptions it is built upon. Considering this, I could give you an example of a COVID-19 mortality early prediction model that was commonly used in the earliest phases of the epidemic. The "Case Fatality Rate" (CFR), which estimates the percentage of confirmed COVID-19 patients that have died, is one such model. Despite not taking into consideration undiagnosed cases or the whole population, this statistic gives an assessment of the disease's severity (Gershengorn et al 2021). To simulate the transmission of the virus and calculate the possible number of infections and fatalities, other prediction models are used, such as the "Susceptible-Exposed-Infectious-Removed" (SEIR) model. These models take into account a variety of variables, including healthcare availability, infection rates, transmission patterns, and population demographics. But it’s important to note that when new information becomes available and the situation changes, these models need to be updated often (Maghbooli et al 2020). Get the most recent and precise COVID-19 mortality prediction models, it is crucial to visit reliable sources such government health organizations, academic institutions, and epidemiological specialists. Based on recent advancements and research, these sources may provide you the most up-to-date knowledge and perceptions (Kassirian et al 2020). Significant respiratory sickness brought on by COVID-19 may need ventilator assistance in certain individuals. The term "ventilator support" describes the use of artificial ventilation or other methods to supplement or take the place of the body’s natural breathing mechanism (Arvind et al 2021). Patients with COVID-19 whose respiratory symptoms are severe and whose illness is progressing often need ventilator assistance. When given supportive treatment, such as rest, water, and monitoring, many virus-infected people recover from their mild to severe symptoms. Nevertheless, if the system for breathing is seriously damaged, more dangerous conditions such as acute respiratory distress syndrome (ARDS) and respiratory failure may arise (Lu et al 2020).

In these circumstances, ventilator assistance is essential for maintaining normal breathing and supplying oxygenation. When a COVID-19 patient is critically sick, mechanical ventilation may assist provide oxygen to the lungs and eliminate carbon dioxide from the body, easing their respiratory distress. Based on the patient’s health and the medical team’s
evaluation, several ventilation techniques, such as positive pressure ventilation, may be used (Goyal et al 2021). It's important to highlight that the choice to begin ventilator support is decided by medical specialists based on evaluations of each individual patient, taking into consideration elements like saturation with oxygen levels, breathing rate, results from imaging, and other clinical indications. When choosing a course of therapy for COVID-19 patients who need ventilator support, the availability of intensive care unit (ICU) resources and ventilators is particularly important (Barbosa et al 2023). It's essential to note that a number of variables, including age, underlying medical problems, and access to treatment, might affect the fatality rate of COVID-19. Ventilators assistance is a crucial procedure that may greatly increase the odds of life for patients who are critically sick, but it does not ensure success. The overall death rates related to COVID-19 are continuously updated as new information becomes available and as medical procedures and methods of treatment advance (Núñez et al 2021).

Early prediction models for COVID-19 mortality and the requirement for ventilator support have been devised to help medical facilities and states anticipate the possible impact of the global epidemic and allocate assets effectively. These models are designed to pinpoint risk-raising variables and provide suggestions for lowering mortality and the need for ventilator assistance (Bonadia et al 2020). Age and current health issues Age and pre-existing conditions including cardiovascular disease, diabetes, obesity, and respiratory diseases increase the risk of COVID-19-related severe illness and death. Early theories highlighted the need of locating and safeguarding these vulnerable people (Wollenstein-Betch et al 2020). Early-stage prediction models for hospital capacity have emphasized the need of providing enough hospital capacity, including ICU beds, ventilators, and medical staff. The need for ventilator support may be met and death rates can be lowered by maintaining enough resources (Twe et al 2022). Early detection and action: Prompt detection of severe cases and early action may enhance results. It is possible to take prompt action, such as providing oxygen treatment and ventilator support, if necessary, by identifying those who are more likely to have severe sickness and carefully monitoring their health (Calligaro et al 2020). The chance of a fatal COVID-19 infection and severe sickness may be significantly decreased with vaccination. In order to safeguard vulnerable people and lessen the overall strain on healthcare systems, early prediction models have underlined the significance of vaccination efforts (Dayan et al 2021). To examine reliable sources including government health organizations, research institutes, and expert consensus statements from epidemiologists and healthcare specialists in order to acquire the most current and accurate prediction models and recommendations. These resources may provide precise recommendations that are adapted to the present level of awareness and comprehension of COVID-19 (Carrasco-Sánchez et al 2021).

The COVID-19 patients that had CXR in the ED between March 1 and March 13, 2020, are prospectively included. Accumulation on the CXR and ground-glass transparency (GGOs) was each assessed independently by two assessors. Two different scoring methods were used. The reliability between observers was assessed using weighted Cohen’s kappa or the intraclass correlation coefficient (ICC). By using logistic or Poisson regression analysis, indicators of mortality and respiratory assistance were found (Balbi et al 2021). The study analyzes a database of blood samples from 485 infected individuals in the Wuhan region of China to identify important prognostic markers of disease mortality. The decision-making procedure described in the article may be used to quickly identify high-risk patients, provide them priority treatment, and perhaps reduce the mortality rate (Yan et al 2020). Sub analysis of a hypothetical, global database of critically ill COVID-19 patients. Patients were subcategorized into high-flow oxygen therapy (HFNC), noninvasive ventilation with positive pressure (NIV), and early IMV in accordance with the breathing assistance technique employed on the day of admission to the ICU. To establish comparability across groups, propensity score matching was used (Wendel Garcia et al 2021). The goal of the research is to summarize P-SIL etiology and the role of passive oxygen therapy in COVID-19 pneumonia. Debates are had on the esophagus volume swings, breathing structure, vascular dysfunction, elastomeric properties, homogenization of the lungs, and effort while breathing (Battaglini et al 2021). The goal of the study was to determine whether there was a connection between blood sodium levels while a patient was hospitalized and significant clinical outcomes like mortality, the requirement for advanced respiratory support, and acute kidney damage (AKI), in addition to whether serum sodium could be used to gauge a state of inflammation in COVID-19 (Tzoulis et al 2021). The decision models for a random forest method, Least Absolute Shrinkage and Selection Operator (LASSO), and reverse selection were combined to provide a projected risk score. Death, the need for or tracheal intubation, and the necessity for minimally invasive breathing were all considered severe outcomes. The research comprised 610 individuals in all, 313 of whom had a catastrophic result (Ageno et al 2021). In the research, they included successive hospitalized with PCR-confirmed COVID-19 who received treatment to Newcastle upon Tyne Hospitals and University College London Hospitals in the UK during the first wave of community-acquired infection. The demographics, research results, and medical condition were documented from the day of hospitalization until death throughout a maximum monitoring period of 28 days (Manson et al 2020).

2. Materials and Methods

The study’s primary goals were to identify patients who might benefit from ventilator assistance and to forecast patient death. For patients with COVID-19 who were hospitalized, the first instance indicated death, whereas cases 2 and 3 indicated ventilator support. The whole feature collection was used to perform three sets of experiments for each case, and the EAI
feature importance approach was used to choose a selected group of characteristics for each trial. In the third trial, instances 2 and 3 only used CXR features, whereas case 1 was completed using the co mortality component. Figure 1 depicts the flow of the overall methodology.

Figure 1 Flow of the proposed methodology.

2.1. Dataset

The study examined retrospectively data from individuals who had COVID-19 and got hospitalized in the Kingdom of Saudi Arabia (KSA). Data from clinical and laboratory testing, as well as CXR findings, are included in the collection along with information on 5739 patients. Additionally, the dataset includes two objective attributes: ventilator support and patient outcome. The eligibility requirements for the patient sample used in this analysis are the same as those used to compile the mortality and ventilator support datasets. The data set contains details on the demographics, testing results from “complete blood counts”, “mean corpuscular hemoglobin concentration”, “mean corpuscular hemoglobin”, “mean corpuscular volume”, “mean platelet volume”, “red blood cells”, “platelet count, red cell distribution width”, “white blood cells”, radiological results, and co-morbidity. The only numerical qualities are age and all CBC characteristics; all other variables are qualitative. Figure 1 depicts the 12 zones in which the CXRs are marked. The junction (C), upper (A), and lower zones (B) are the first divisions of the CXR. Then, twelve more zones are created from these zones to represent the locations where a radiologist assigns seriousness ratings. There are three potential values (0–2) for the zone characteristics that represent the degree of ground glass opacity (GGO). GGO is not present when the value is zero. In all, there are 2 class characteristics and 35 predictors in the dataset. Additionally, the age feature was separated using equal-width binning for the dataset's an exploratory examination. The dataset's lowest and highest patient ages respectively were 19 and 107.

Figure 2 Zone segmentation and annotation for chest X-rays.
2.2. Preprocessing using Z-score normalization

Z-score normalization, usually referred to as standardization, is a method frequently used in predictive modeling to standardize numerical data. By dividing by the standard deviation and removing the mean from the data, it turns the data into a distribution with a mean of zero and a standard deviation of one. This normalization approach may be used to verify that the characteristics in a COVID-19 mortality prediction model are on a comparable scale.

Obtain the information you want for your prediction model, including the target variable (COVID-19 mortality) and any pertinent characteristics (such as age, co morbidities, vital signs, etc.).

Data should be normalized for each characteristic before being divided by the standard deviation and the mean. The z-score normalization formula (1) is as follows:

\[ Z = \frac{x - \mu}{\sigma} \quad (1) \]

Where:
The normalized value (z-score) is called z.
X is the feature’s initial value.
\( \mu \) is the feature’s mean.
\( \sigma \) is the feature’s standard deviation.

Use this formula to convert each feature in your dataset into its corresponding z-score for each data point.

The values obtained after normalizing the features using z-score normalization may be used as inputs for your COVID-19 mortality prediction model. Make careful to normalize any additional data that you wish to use for predictions using the same mean and standard deviation figures that you acquired from your training data. It is simpler for the model to learn from the data when the scale discrepancies between the features are removed via Z-score normalization. Additionally, it avoids the learning process from being dominated by characteristics with greater sizes. To the characteristics, not the target variable, normalization should be done. The COVID-19 mortality goal variable should be left normalized and in its original state. It is important to remember that there are several normalizing methods available, such as min-max scaling, which scales the data within a certain range. Your dataset’s unique needs and features, as well as the selected predictive modeling technique, will influence the normalization approach you choose.

2.3. Feature extraction using Linear discriminate analysis

A statistical technique known as linear discriminate analysis (LDA) is often used to resolve issues involving pattern recognition. LDA was first used for face recognition. Fisher’s Linear Discriminate is the name of this approach. To extract crucial information from data and minimize data dimensionality, LDA is becoming increasingly often utilized. To divide the data into two or more object classes or groups, this approach essentially seeks for linear combinations of properties that represent the primary characteristics. The distribution of fresh data will be more diffused after it has undergone LDA processing, which will ultimately improve recognition success. Following that, both linear and non-linear classification techniques may be used to reprocess the new data combination that was produced. The primary properties of the Caesar dataset are determined using LDA in this study, and the researcher also attempts to decrease the dimensions from the available data if at all feasible. This is done in the hopes that the outcomes of the dimensional reduction would be correctly categorized using SVM. The LDA algorithm is shown below. To better understand them; let’s first create the covariance matrix in class SW and the interclass covariance matrix SB, each of which is shown by the following for formula (2).

\[ T_X = \sum_{j=1}^{d} \sum_{w_i \in W_j} (w_l - \mu_j) (w_l - \mu_j)^S \quad (2) \]

\[ T_X = \sum_{j=1}^{d} M_j (w_l - \mu_j) (w_l - \mu_j)^S \quad (3) \]

Where \( X_k \) denotes data in the k-position, C is the whole class, \( N_i \) denotes the quantity of data in the class, \( \mu \) denotes the average across all data, and \( \mu_j \) denotes the average data in class i.

A matrix of covariance across classes is then maximized in order to reduce the covariance matrix inside the class. The eigenvector may then be found, allowing us to maximize the ratio of (4):

\[ \frac{\text{det}(UT_A U^S)}{\text{det}(UT_X U^S)} \quad (4) \]

\[ T_A = \lambda T_X U \quad (5) \]

\[ \text{cov} = T_A (T_X)^{-1} \quad (6) \]
The eigenvector utilized is $c^{-1}$ since the LDA technique covered in this study is the LDA method with two classes. The biggest eigenvalue of the covariance matrix may be used to determine the eigenvector value.

Once the eigenvector is known, the LDA feature may be calculated by doing the following formula:

$$E_w = \sum_{j=1}^{l} (w_j - \mu)^T \times U$$  \hspace{1cm} (7)

2.3. Predictive model using hybrid spider monkey optimized eXtreme gradient boosting (HSMO-XGB)

2.3.1. Spider Monkey Optimization

A more modern optimization technique based on swarm intelligence is the hybrid spider monkey optimization (SMO) algorithm. Distances calculated according to between possible solutions serve as the foundation for the update equations. The technique has been used widely to resolve challenging optimization issues. A fuzzy rule foundation was designed and optimized using SMO. The SMO method optimizes the amplitudes of each element and the positions of each element in the extended sparse sub array to minimize the side lobe levels of the whole array while adhering to a number of real-world limitations. Built SM-Rule Miner to do rule extraction on diabetic data. The array factor of a linear antennas array was shaped antenna patches that has been perfectly designed for use in wireless networks.

2.3.2. Extreme gradient boosting

The Extreme Gradient Boosting (XGBoost) technique is a regression tree with decision-tree-like decision-making capabilities. It supports both classification and regression. This approach is a scalable and effective variation of the gradient boosting machine (GBM), which has been extensively used in computer vision, data mining, and other areas. Recently, XGBoost has primarily advanced in two areas as a sort of gradient boosting machine: speeding up the tree building and suggesting a new distributed technique for tree searching. The heart of XGBoost is optimizing the value of the goal function. Considering the dataset $D = \{\{x_i, y_i\}\}$ represents the xi tumor’s gene expression profile and the related yi binary label (early stage or late stage) is used to identify the tumor. Equation is the optimization objective function for the XGBoost model with $K$ decision trees.

$$\hat{z}_j = \sum_{s=1}^{l} e_s(x_j) e_s e_E$$  \hspace{1cm} (8)

Where each $f_k$ represents a separate tree with leaf scores and $F$ is the space of the regression tree. In formula (9), the loss function is shown.

$$K(e_k) = \sum k (\hat{z}_j, z_j) + \sum \Omega e_k$$  \hspace{1cm} (9)

Differentiable loss function $l$ in the first term assesses the discrepancy between the output $\hat{y}_i$ that was anticipated and the actual output $y_i$.

$$\hat{z}_j^{(s)} = \hat{z}_j^{(s-1)} + e_s(x_j)$$  \hspace{1cm} (10)

$$\Omega(e) = \gamma S + \frac{1}{2}||x||^2$$  \hspace{1cm} (11)

$W$ is the score on each leaf, and $T$ is the total number of leaf nodes. Therefore, we might conclude that

$$S(e_s) \approx \sum_{i=1}^{S} \left[ (\sum_{j \in E_i} h_j) x_i + \frac{1}{2} (\sum_{j \in E_i} g_j + \lambda) x^2_i \right] + \gamma S$$  \hspace{1cm} (12)

Where $gi$ and hi are the loss function’s first-order and second-order gradient statistics, respectively.

2.4. Assessment measures

A comparison of “accuracy, sensitivity, specificity, Youden index, and area under the curve” (AUC) was used to assess the proposed model’s efficacy. The effectiveness of the categorization algorithms may be examined using a variety of assessment metrics. Some of these metrics, including accuracy, precision, and recall, are significantly impacted by class distribution. Consequently, the uneven class distribution, including recall, accuracy, and precision. The imbalanced class issue thus employs metrics like balanced accuracy, AUC, and Youden index. Sensitivity (SN), which is generated using the formula below, is used to reflect the model’s genuine positive rate.

$$SN = \frac{\text{sum(positive class samples in the dataset)}}{\text{sum(correctly predicted positive class)}}$$  \hspace{1cm} (13)
The following formula is used to compute specificity, which stands for the real negative rate:

$$SP = \frac{\text{sum(negative class samples in the dataset)}}{\text{sum(correctly predicted Negative class samples)}}$$  \hspace{1cm} (14)

Equations (13) and (14) demonstrate that the following metrics are unaffected by class distribution.

The mean of the sensitivity and specificity is the same for balanced accuracy:

$$\text{Balanced Accuracy} = \frac{\text{SP} + \text{SN}}{2}$$  \hspace{1cm} (15)

One of the metrics used to assess how well the diagnostic test works is the Youden index (YI). It rates the diagnostic test's capacity for discrimination. The following equation is used to calculate it:

$$\text{Youden Index} = \text{SN} + \text{SP} - 1$$  \hspace{1cm} (16)

The YI values range from 0 to 1. A result that is closer to one demonstrates the test's relevance, while a score that is lower indicates insufficient diagnostic ability.

AUC is used to confirm the model’s ability to discriminate further. AUC values range from 0 to 1, and like the Youden index, values closer to 0 indicate poor performance while values closer to 1 indicate significant effectiveness of the model.

3. Results

In this section, the results for all three situations that indicate that COVID-19 patients will die and need ventilator assistance are discussed. Additionally, the dataset is unbalanced, which is why data-sampling strategies like SMOTE with oversampling and SMOTE with under sampling were used. A cross-validation with process was used to split the data among all trials, with k set to a value of 10. After that, training and validation sets were created from the training parts. The validation set was used throughout the parameter tweaking procedure. Figure 3 depicts the results of the experimentation using the recommended model for Case 1 in order to establish when the patient would be either deceased. Similar responsiveness is accomplished utilizing the full-and chose highlight sets with the underlying data set without the utilization of any information testing methods. After SMOTE under sampling, the whole feature set and dataset produced the results with the maximum sensitivity. However, the entire feature set with oversampling and SMOTE produced the best results for other metrics like specificity, balanced accuracy, the Youden index, and AUC.

A comparable AUC for SMOTE surpassing was achieved using the whole feature set as well as the original dataset with the specified features. Using specific characteristics and SMOTE oversampling, the best overall AUC was found.

![Graph A](image1.png)

![Graph B](image2.png)

![Graph C](image3.png)

**Figure 3** The outcome of the suggested mortality prediction model.
However, utilizing the co morbidity variables and the original dataset, the maximum AUC attained in the suggested research was 0.875, and the balanced accuracy was 0.904. We discovered that utilizing co morbidity features, a comparable AUC was obtained after oversampling the sample. Additionally, studies were carried out to determine whether patients need ventilator assistance. The tests were at first done for the numerous classes, i.e., to lay out when patients required a “mechanical ventilator”, a “painless ventilator”, utilizing each quality, chose highlights, and CXR highlights. The use of CXR facial appearance was made possible by the author’s discovery that CXR traits could be used to forecast whether COVID-19 patients would need ventilator assistance during a foundation experiment. The effectiveness of the suggested DL model employing various feature sets is shown in Figure 4.

![Figure 4](https://www.malque.pub/ojs/index.php/msj)

**Figure 4** The outcome of the suggested ventilator-support prediction model (Multiclass).

![Figure 5](https://www.malque.pub/ojs/index.php/msj)

**Figure 5** Results of the ventilator-support prediction model suggested.
The test results are listed in the table. Similar to scenario 1, the model in this instance performed best when all features were enabled. It shows the importance of each characteristic in determining whether individuals need ventilator assistance. SMOTE with the under-sampling dataset set, however, produced the best outcomes in this instance for all assessment metrics. Both the full-feature set and the chosen feature set produced a comparable AUC for the mortality prediction (case 1) task. However, the model AUC utilizing complete and chosen characteristics in this instance differed significantly. In addition, we discovered that the oversampling method outperformed the original and under sampled datasets in terms of performance with CXR features. To determine which patients would need ventilation, trials for the binary class were carried out. Figure 5 uses the test set to display the outcome of the suggested model. Similar to example 2, the best results were also obtained for a single class using full features and under sampled data. The model's performance was marginally enhanced after switching from a multiclass to a binary class. However, there were substantial differences between the results of the full-feature dataset and the other datasets, such as the original and over- and under-sampling.

They discovered that XGB and the dataset with a low sample size produced the best results for determining whether the patient would require ventilator support. With an AUC of 0.904 and balanced accuracy of 0.875, the proposed research outperformed a foundational study when applied to the initial dataset. All findings show that the suggested model is significant in each of the three scenarios.

4. Discussion

It is essential to develop an electronic model that is able to identify patients who are at risk for mortality and provide assistance for ventilators due to the severe clinical signs of Coronavirus and, occasionally, its rapid breakdown in moderate-stage patients. Additionally, a model that can provide healthcare practitioners a trustworthy explanation must be made available. As a result, to predict mortality and ventilator support in the planned research, the DL model was combined with the EAI (Vila Corcoles et al 2021). Multiple studies have examined the use of laboratory results, signs and symptoms, data on demographics, and radiological results to make predictions. As a result, the suggested research made use of demographic, clinical, co morbidities, and CXR zone characteristics. Age and gender were shown to be important elements among the demographic characteristics. To predict mortality and ventilator support, however, radiomics characteristics and two demographic factors (age and gender) were utilized. Experienced radiologists who gave the radiomic ratings discovered that radiomic characteristics significantly improved the efficiency of the algorithms. Included pharmaceutical data and discovered that individuals on medication for pneumonia and respiratory disorders were more likely to use a ventilator. On the other hand, it was discovered that several of the patients who had received a COVID-19-negative diagnosis through RT-PCR test had pneumonia upon CXR examination. Similar to this, several of the patients’ CXR results showed no significant indications, but the RT-PCR test anticipated that they would be positive. But they discovered that CXR properties can only be utilized for prediction if additional variables, such as SpO2, PaO2, and certain other clinical factors, are also present. On the other hand, use demographic, co morbidity, and symptom data to forecast death (Wendel Garcia et al 2021). Despite this, the research produced important findings; however, they were missing certain important CXR and lab test data from the dataset.

Additionally, three co morbidities (heart issues, diabetes, and hypertension) were looked at as critical aspects. However, it was shown that although several lab tests were a significant factor in predicting death, the top attributes to do so were shortness of breath and further oxygen treatment. Even so, the present research has shown notable findings, although there is always opportunity for advancement. The dataset used for the research came from a single center and nation (Karthikeyan et al 2021). Several clinical parameters that have been found to be crucial are also absent from the dataset, including CPR, D-dimer, heartbeat, SpO2, and PaO2. To further confirm the effectiveness of the model, it must be examined using the multicenter dataset while also taking into consideration other elements that have been shown to be pertinent in earlier studies.

5. Conclusions

This study examined whether Coronavirus patients required ventilator support and whether the DL model could predict mortality. The dataset contains information on the demographics, test outcomes, co-morbidities, and CXR of the COVID-19 patients. To solve the issue of data imbalance, both under- and over-sampling were done using the SMOTE data-sampling method. The EAI feature significance method was used to choose the features. Utilizing the hybrid spider monkey optimized, the extreme gradient boosting was optimized. For the purpose of predicting mortality and ventilator support, many sets of experiments were conducted utilizing CXR results and complete features, chosen features, and co morbidity characteristics alone. The suggested model may be used, in particular, as a tool to help clinicians forecast patients who are at risk and to help hospitals manage and organize their resources efficiently. On the other hand, this research may be expanded to look at performance utilizing the multicenter and multicounty dataset. It’s also necessary to take into account several of the key test investigation outcomes and COVID-19 immunizations.

References

https://www.malque.pub/ojs/index.php/msj


Barbosa HC, Martins MAP, Jesus JCD, Meira KC, Passaglia LG, Sacioto MF, Marcolino MS (2023) Myocardial Injury and Prognosis in Hospitalized COVID-19 Patients in Brazil Results From The Brazilian COVID-19 Registry. Arquivos Brasileiros de Cardiologia 120. DOI: 10.36660/abc.20220151


